

Deep Learning

Assignment 5: Spoken Digits Recognition

Instructor: Dr. Mehrandezh
Student: Marzieh Zamani

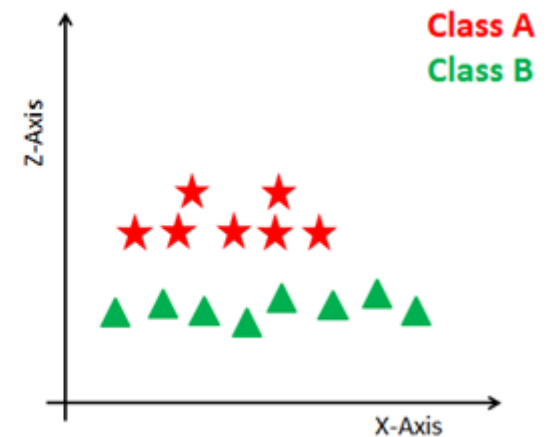
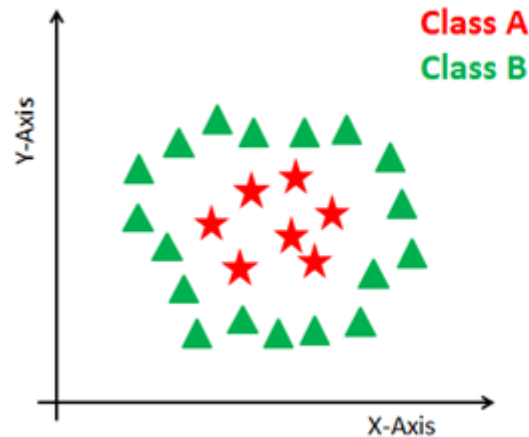
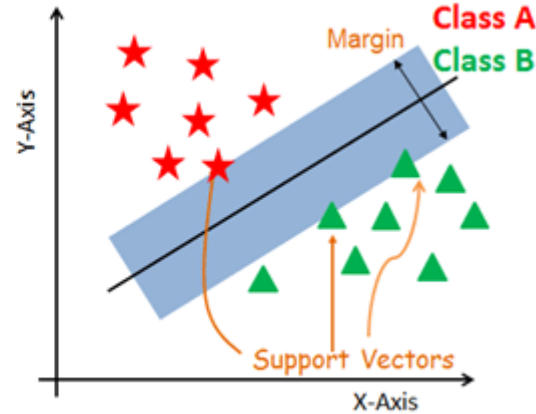
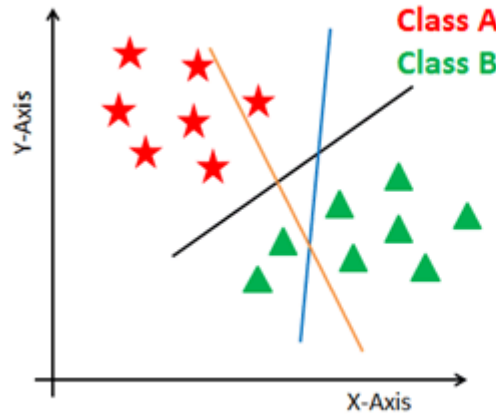
Spoken Digit Recognition on FSDD Dataset

- ▶ FSDD Dataset (2 native and 2 non-native speakers)
 - Training: 160 recordings per digit
 - Testing: 40 recordings per digit
- ▶ Student Dataset (1 non-native speaker)
 - Evaluation: 5 recordings per digit
- ▶ Sampling Frequency: 8000
- ▶ Mono

Classifiers

- ▶ Support Vector Machine (SVM)
- ▶ Long Short-Term Memory (LSTM)
- ▶ LSTM optimized by Bayesian Optimization
- ▶ Deep Convolutional Neural Network (DCNN) with Mel-frequency Spectrograms

Support Vector Machines (SVM) Classifier



SVM Classifier

- ▶ **Specification:** quadratic polynomial kernel
- ▶ **Feature vector:** $N \times 321 \times 1$
- ▶ **Training dataset:** FSDD ($N=160 \times 10$)
- ▶ **Testing dataset:** FSDD ($N=40 \times 10$)
- ▶ **Evaluation dataset:** Student ($N=5 \times 10$)

SVM Training & Testing on FSDD Dataset => 98%

SVM Classification | FSDD Testing Dataset

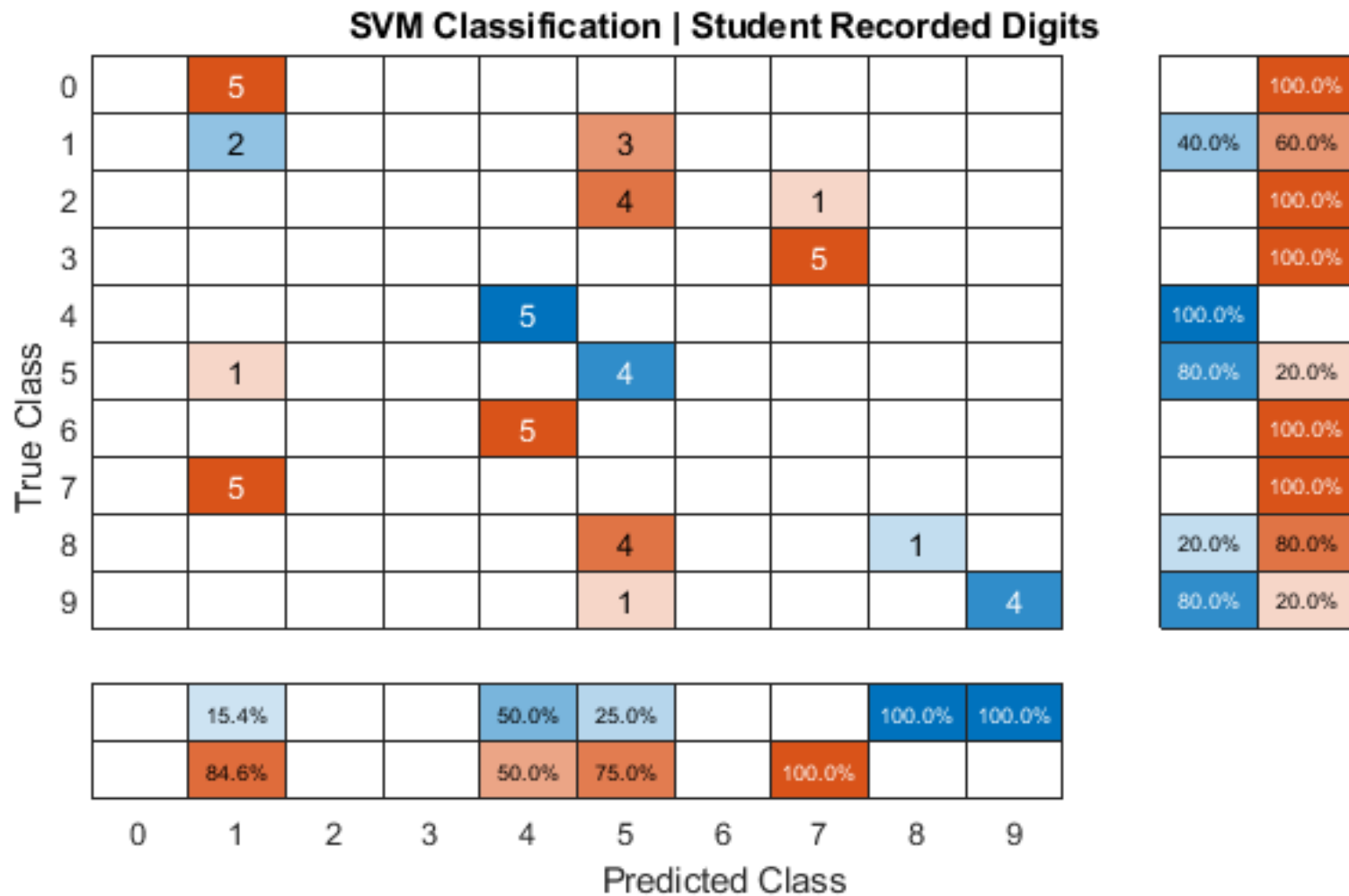
True Class	0	39					1						97.5%	2.5%
	1		40										100.0%	
	2	1		39									97.5%	2.5%
	3			2	38								95.0%	5.0%
	4		1			39							97.5%	2.5%
	5					1	38		1				95.0%	5.0%
	6				2			38					95.0%	5.0%
	7								40				100.0%	
	8									40			100.0%	
	9										40		100.0%	

97.5%	97.6%	95.1%	95.0%	97.5%	97.4%	100.0%	97.6%	100.0%	100.0%
2.5%	2.4%	4.9%	5.0%	2.5%	2.6%		2.4%		
0	1	2	3	4	5	6	7	8	9

Predicted Class

SVM Evaluation on Student Dataset

=> 32%



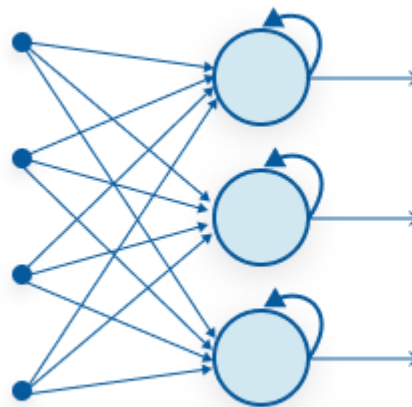
SVM Remarks

- ▶ **Testing dataset:** FSDD ($N=40 \times 10$) \Rightarrow test accuracy= 98%
- ▶ **Evaluation dataset:** Student ($N=5 \times 10$) \Rightarrow test accuracy= 32%
- ▶ SVM classifies FSDD dataset with an excellent accuracy of 98%; however, its classification accuracy for student spoken digits is significantly lower, 32%.
- ▶ It is worth mentioning that the classification accuracy of different digits are very distinct, ranging from 0% to 100%.

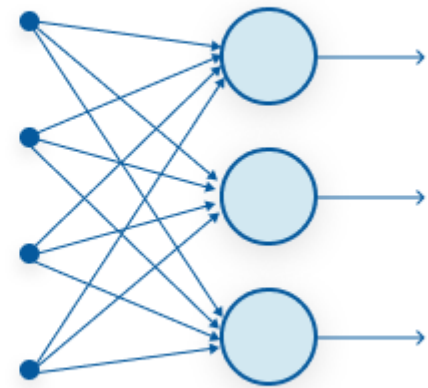
Feed-forward vs. Recurrent NN

- ▶ Feed-forward neural networks pass the data forward from input to output
- ▶ RNN has a feedback loop where data can be fed back into the input

Recurrent Neural Network structure



Recurrent Neural Network



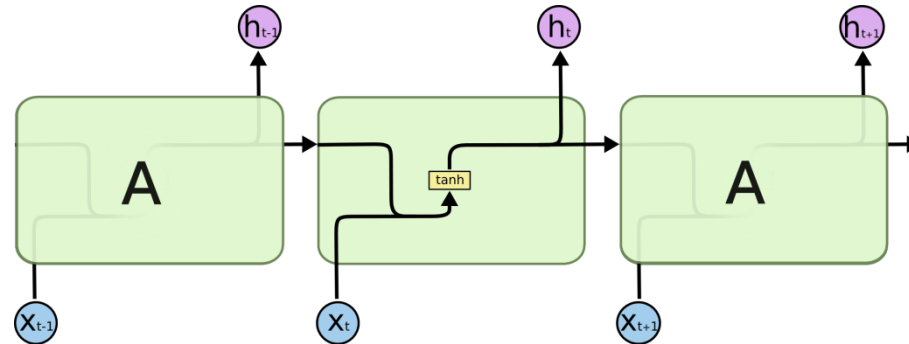
Feed-Forward Neural Network

Long Short-Term Memory (LSTM) Classifier

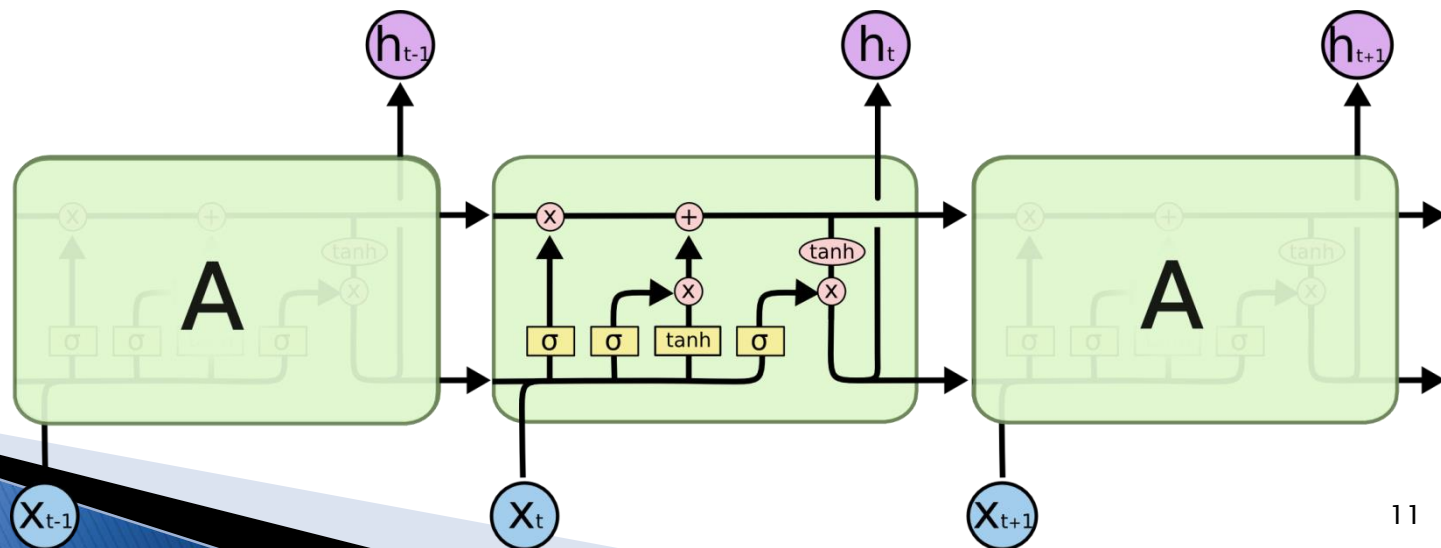
- ▶ Designed to remembering information for long periods of time;
- ▶ The cell state is kind of like a conveyor belt of information
- ▶ Gates: optionally let information to cell state.

Long Short-Term Memory (LSTM) Classifier

- ▶ The repeating module in a standard RNN contains a single layer.



- ▶ The repeating module in an LSTM contains four interacting layers.



LSTM Architecture & Parameters

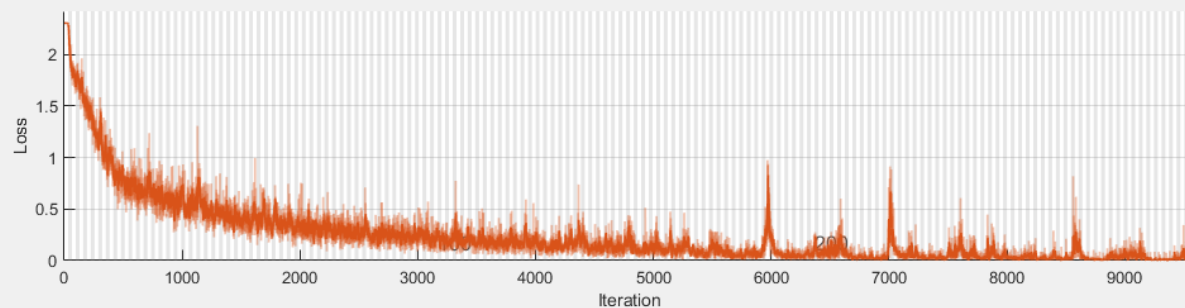
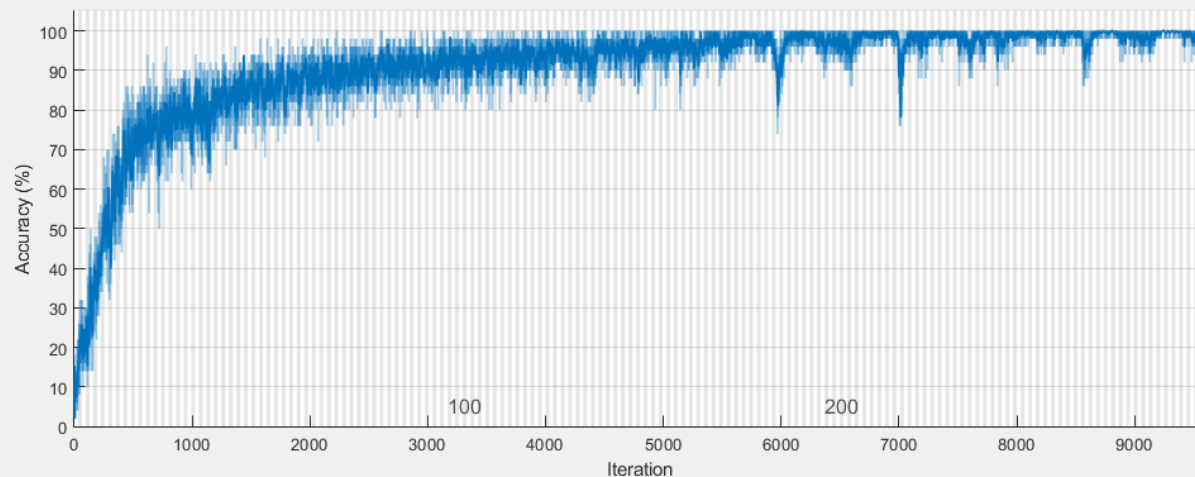
- ▶ 512 hidden units
- ▶ Initial Learning Rate: 0.0001

No	Layer Type	Output Size	Details
1.	INPUT	321	-
2.	LSTM	512	Input Weights: 2048x321 Recurrent Weights: 2048x512 Bias: 2048x1
3.	FC	10	Weights: 10x512 Bias: 10x1
4.	SOFTMAX	10	-
5.	Class Output	1	-

LSTM Training on FSDD Dataset

Training Progress (24-Jun-2020 12:52:02)

Training Progress (24-Jun-2020 12:52:02)



Results

Validation accuracy: N/A
Training finished: Reached final iteration

Training Time

Start time: 24-Jun-2020 12:52:02
Elapsed time: 391 min 56 sec

Training Cycle

Epoch: 300 of 300
Iteration: 9600 of 9600
Iterations per epoch: 32
Maximum iterations: 9600

Validation

Frequency: N/A

Other Information

Hardware resource: Single CPU
Learning rate schedule: Constant
Learning rate: 0.0001

Accuracy

— Training (smoothed)
— Training
— Validation

Loss

— Training (smoothed)
— Training
— Validation

= > 94%

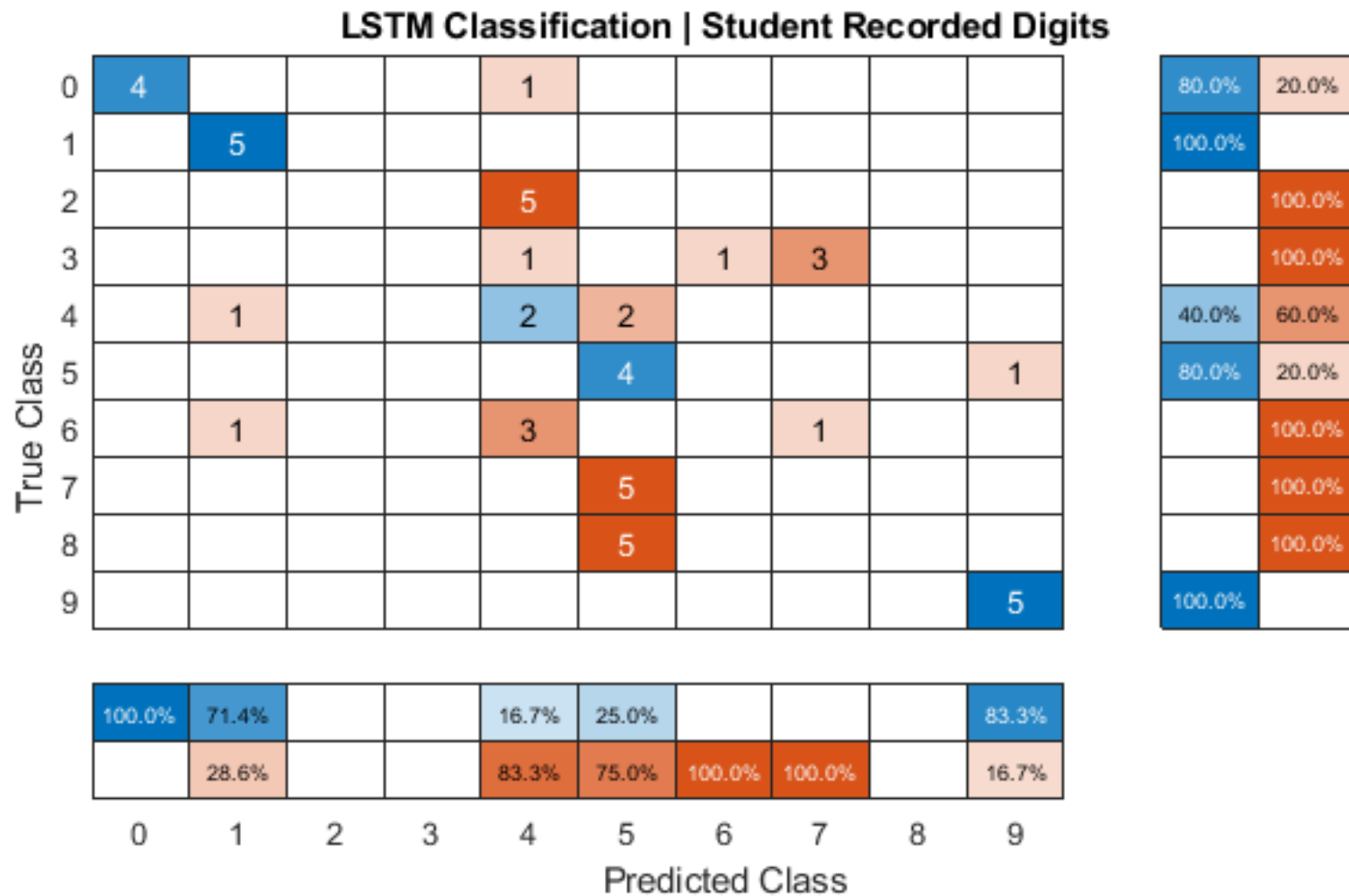
LSTM Classification | FSDD Testing Dataset

True Class	0	38		1						1	95.0%	5.0%
	1		37				1		1		92.5%	7.5%
	2	1		38		1					95.0%	5.0%
	3				39					1	97.5%	2.5%
	4					40					100.0%	
	5						40				100.0%	
	6			1	1			37	1		92.5%	7.5%
	7							2	38		95.0%	5.0%
	8							1		39	97.5%	2.5%
	9		1				1		2		36	90.0%

97.4%	97.4%	95.0%	97.5%	97.6%	95.2%	92.5%	90.5%	97.5%	94.7%
2.6%	2.6%	5.0%	2.5%	2.4%	4.8%	7.5%	9.5%	2.5%	5.3%
0	1	2	3	4	5	6	7	8	9

Predicted Class

LSTM Evaluation on Student Spoken Digits => 40%



LSTM Remarks

- ▶ **Testing dataset:** FSDD ($N=40 \times 10$) \Rightarrow test accuracy= 94%
- ▶ **Evaluation dataset:** Student ($N=5 \times 10$) \Rightarrow test accuracy= 40%
- ▶ LSTM classifies FSDD dataset with an excellent accuracy of 94%; however, its classification accuracy for student spoken digits is significantly lower, 40%.
- ▶ It is worth mentioning that the classification accuracy of different digits are very distinct, ranging from 0% to 100%.

LSTM–Optimized by Bayesian Optimization

- ▶ **Optimization parameters and range**
 - Initial Learning Rate: [1e-5, 1e-1]
=> Optimal value = 2.199e-04
 - Number of Hidden Units: [10, 1000]
=> Optimal value = 768
- ▶ **Feature vector:** $N \times 321 \times 1$
- ▶ **Training dataset:** FSDD ($N=160 \times 10$)
- ▶ **Testing accuracy:** FSDD ($N=40 \times 10$)
=> test accuracy: 94% increased to 97.75%
- ▶ **Evaluation dataset:** Student ($N=5 \times 10$)
=> test accuracy: 40% increased to 42%

LSTM-Optimized Testing on FSDD Dataset => 98%

LSTM-Optimized Classification | FSDD Testing Dataset

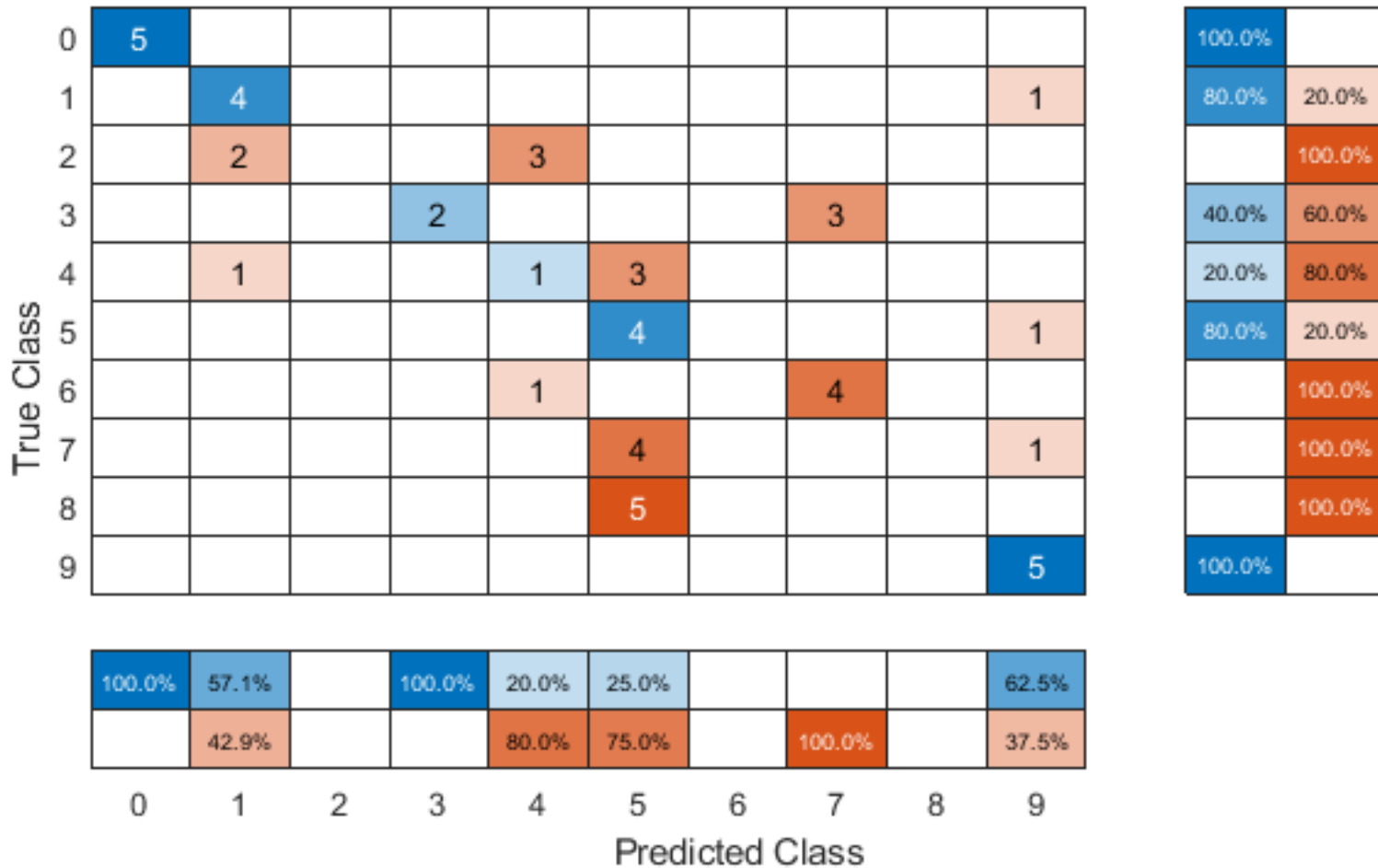
0	38		1						1	95.0%	5.0%
1		37		1	1	1				92.5%	7.5%
2			40							100.0%	
3				40						100.0%	
4		1			39					97.5%	2.5%
5						40				100.0%	
6			1	1			37	1		92.5%	7.5%
7							1	39		97.5%	2.5%
8			1						39	97.5%	2.5%
9		1						1		95.0%	5.0%

100.0%	94.9%	93.0%	95.2%	97.5%	97.6%	97.4%	95.1%	100.0%	97.4%
	5.1%	7.0%	4.8%	2.5%	2.4%	2.6%	4.9%		2.6%
0	1	2	3	4	5	6	7	8	9

Predicted Class

LSTM-Optimized Evaluation on Student Spoken Digits => 42%

LSTM-Optimized Classification | Student Recorded Digits



LSTM-Optimized Remarks

- ▶ **Optimization parameters and range**
 - Initial Learning Rate: $[1e-5, 1e-1]$
=> Optimal value = 2.1988
 - Number of Hidden Units: $[10, 1000]$
=> Optimal value = 768
- ▶ **Testing accuracy: FSDD ($N=40 \times 10$)**
=> test accuracy: 94% increased to 97.75%
- ▶ **Evaluation dataset: Student ($N=5 \times 10$)**
=> test accuracy: 40% increased to 42%

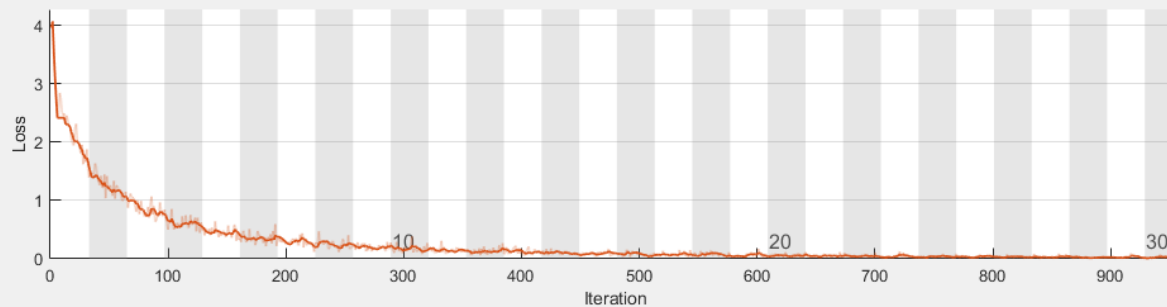
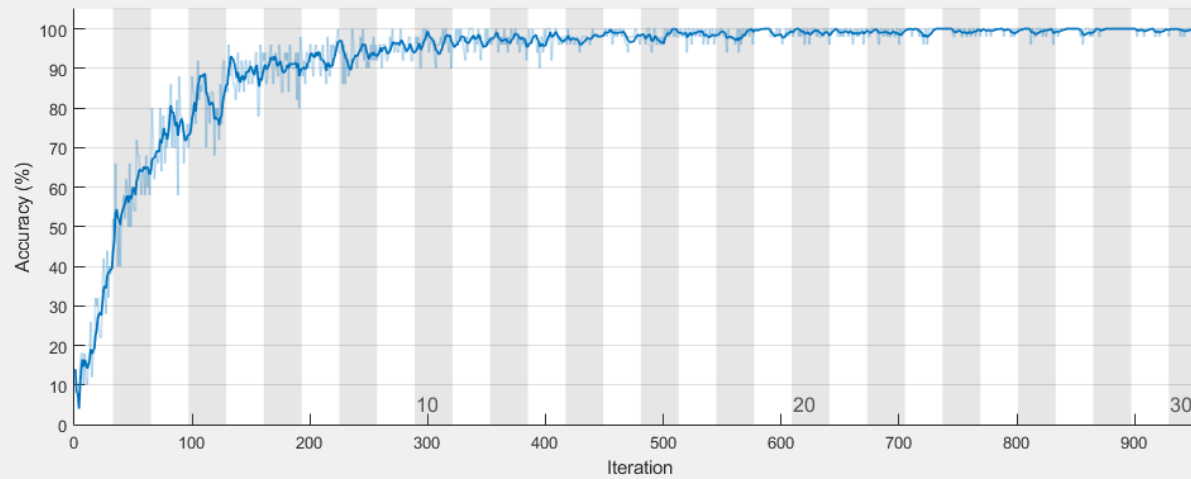
DCNN Architecteur and Parameters

No	Layer Type	Output Size	Filter Size / Stride
1.	INPUT IMAGE	40×81×1	
2.	CONV	40×81×12	5×5;K = 12
3.	BN	40×81×12	
4.	ACT (ReLU)	40×81×12	
5.	POOL	20×41×12	F=3, S=2
6.	CONV	20×41×24	3×3;K = 24
7.	BN	20×41×24	
8.	ACT (ReLU)	20×41×24	
9.	POOL	10×21×24	F=3, S=2
10.	CONV	10×21×48	3×3;K = 48
11.	BN	10×21×48	
12.	ACT (ReLU)	10×21×48	
13.	POOL	5×11×48	F=3, S=2
14.	CONV	5×11×48	3×3;K = 48
15.	BN	5×11×48	
16.	ACT (ReLU)	5×11×48	
17.	CONV	5×11×48	3×3;K = 48
18.	BN	5×11×48	
19.	ACT (ReLU)	5×11×48	
20.	POOL	4×10×48	F=2, S=1
21.	DO	4×10×48	dropoutProb = 0.2
22.	FC	1×1×10	
23.	SOFTMAX	1×1×10	
24.	Class Output	1×1×1	

DCNN Training on FSDD Dataset

Training Progress (25-Jun-2020 13:59:57)

Training Progress (25-Jun-2020 13:59:57)



Results

Validation accuracy: N/A
Training finished: Reached final iteration

Training Time

Start time: 25-Jun-2020 13:59:57
Elapsed time: 7 min 40 sec

Training Cycle

Epoch: 30 of 30
Iteration: 960 of 960
Iterations per epoch: 32
Maximum iterations: 960

Validation

Frequency: N/A

Other Information

Hardware resource: Single CPU
Learning rate schedule: Constant
Learning rate: 0.0001

Accuracy

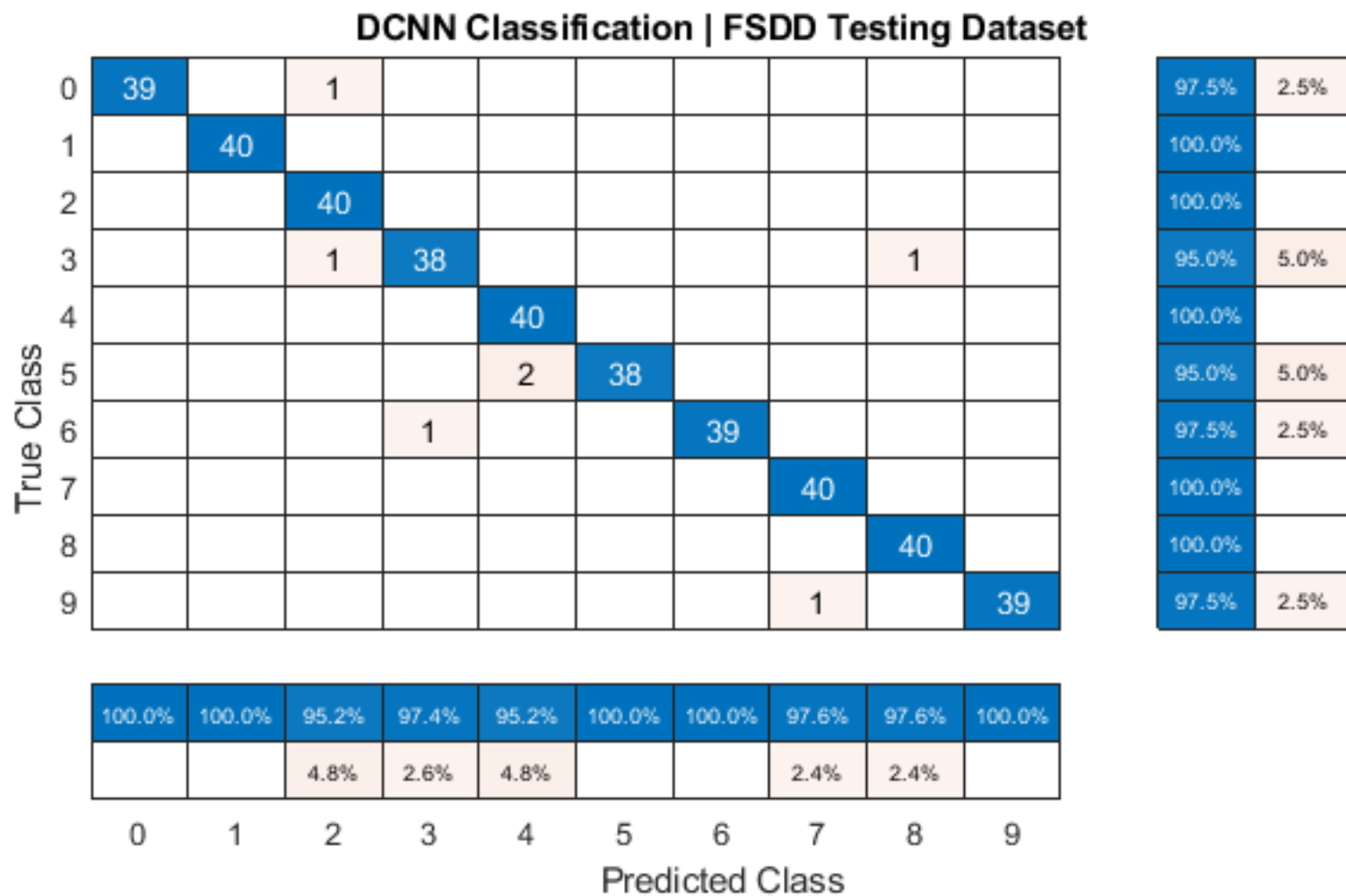
— Training (smoothed)
— Training
— Validation

Loss

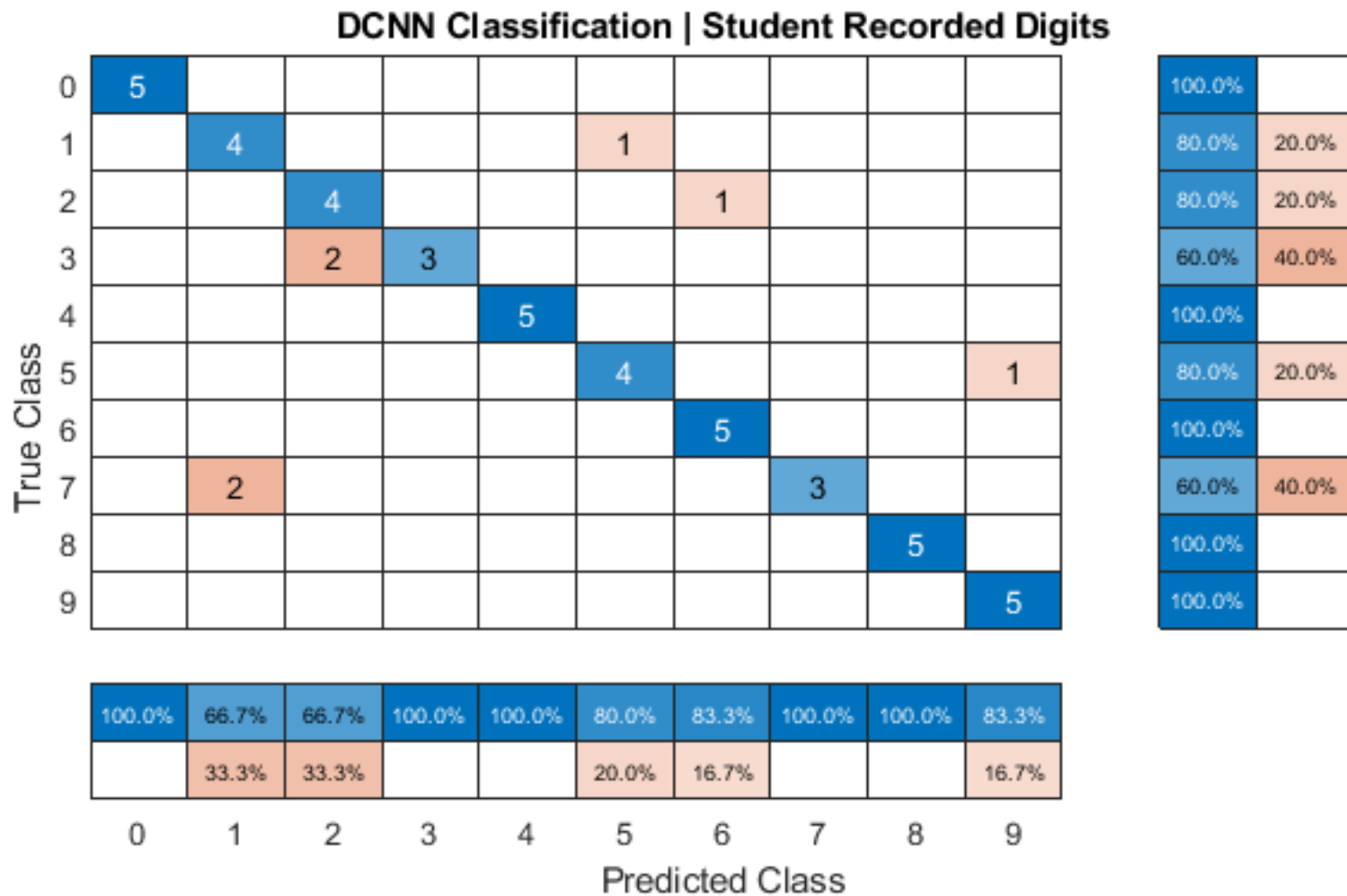
— Training (smoothed)
— Training
— Validation

DCNN Testing on FSDD Dataset

= > 98%



DCNN Evaluation on Student Spoken Digits => 86%



DCNN Evaluation on Student Spoken Digits

- ▶ **Testing dataset:** FSDD ($N=40 \times 10$) \Rightarrow test accuracy= 98%
- ▶ **Evaluation dataset:** Student ($N=5 \times 10$) \Rightarrow test accuracy= 86%
- ▶ DCNN classifies FSDD dataset with a slightly higher accuracy of 98%.
- ▶ Its strength is revealed when it can improve the classification accuracy for student spoken digits from 40% to 86%. This means that not only DCNN is stronger than other networks.

References

1. <https://www.mathworks.com/help/audio/examples/spoken-digit-recognition-with-wavelet-scattering-and-deep-learning.html>
2. Free Spoken Digit Dataset (FSDD), available at <https://github.com/Jakobovski/free-spoken-digit-dataset>
3. <https://www.mathworks.com/videos/series/understanding-wavelets-121287.html>
4. <https://www.datacamp.com/community/tutorials/svm-classification-scikit-learn-python>
5. <https://colah.github.io/posts/2015-08-Understanding-LSTMs/>
6. https://en.wikipedia.org/wiki/Bayesian_optimization#:~:text=Bayesian%20optimization%20is%20a%20sequential,expensive%2Dto%2Devaluate%20functions.

Thanks for your attention



Questions are welcome